Regression models in epidemiology

Tamás Ferenci

25-September-2018

Introduction

Classical and regression models

Regression models in epidemiology

Model diagnostics and -specification: the detective work

Results: using the obtained model

Further questions

Introduction

What is epidemiology?

(An absolutely unofficial definition)

- Who
- When
- Where
- Why

is sick?

Our case study

- Understanding the epidemiology of amputations due to peripheral vascular disease in Hungary
- Using financial/administrative database (covering practically everyone from 2004)
- Age, sex, postal code, procedure/diagnosis date and procedure/diagnosis code is known
- Aims, more specifically: understanding how the disease affects different genders, ages, people with certain comorbities, what are the spatial disparities, what are the time trends (long-term, seasonal, etc.) etc.

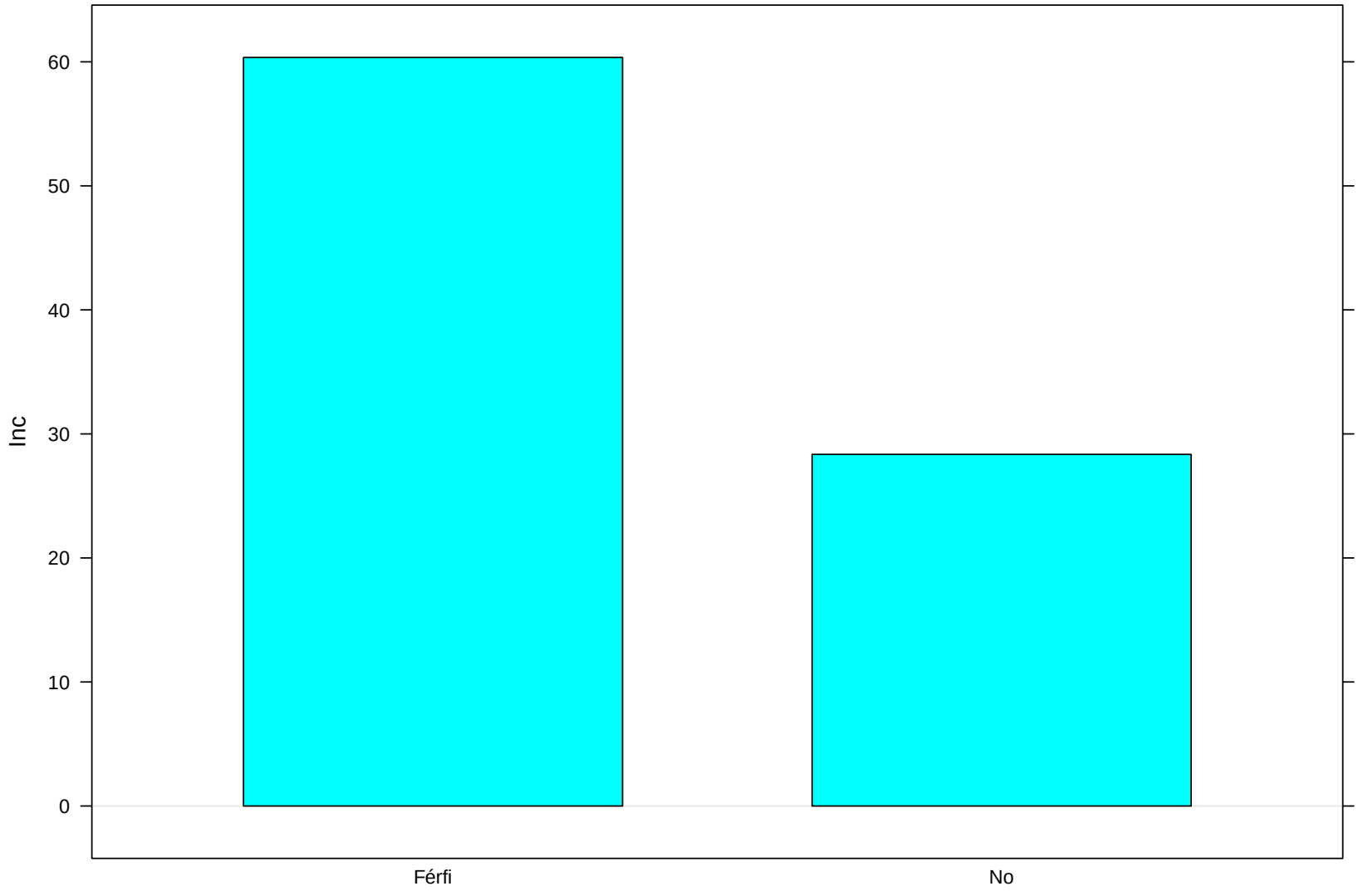
Motivation


Emberi Konzultáció az Egészségügyről
Magyarországon háromszoros az esélyed arra, hogy érszűkület miatt levágják a lábad.*
Legyen az egészségügy Nemzeti Ügy!
Írd alá: reszasz.hu/peticio
*Az Európai Uniós átlaghoz viszonyítva. Forrás: Eur J Vasc Endovasc Surg (2015)
RESZASZ
MUNKÁSTANÁCSOK

Classical and regression models

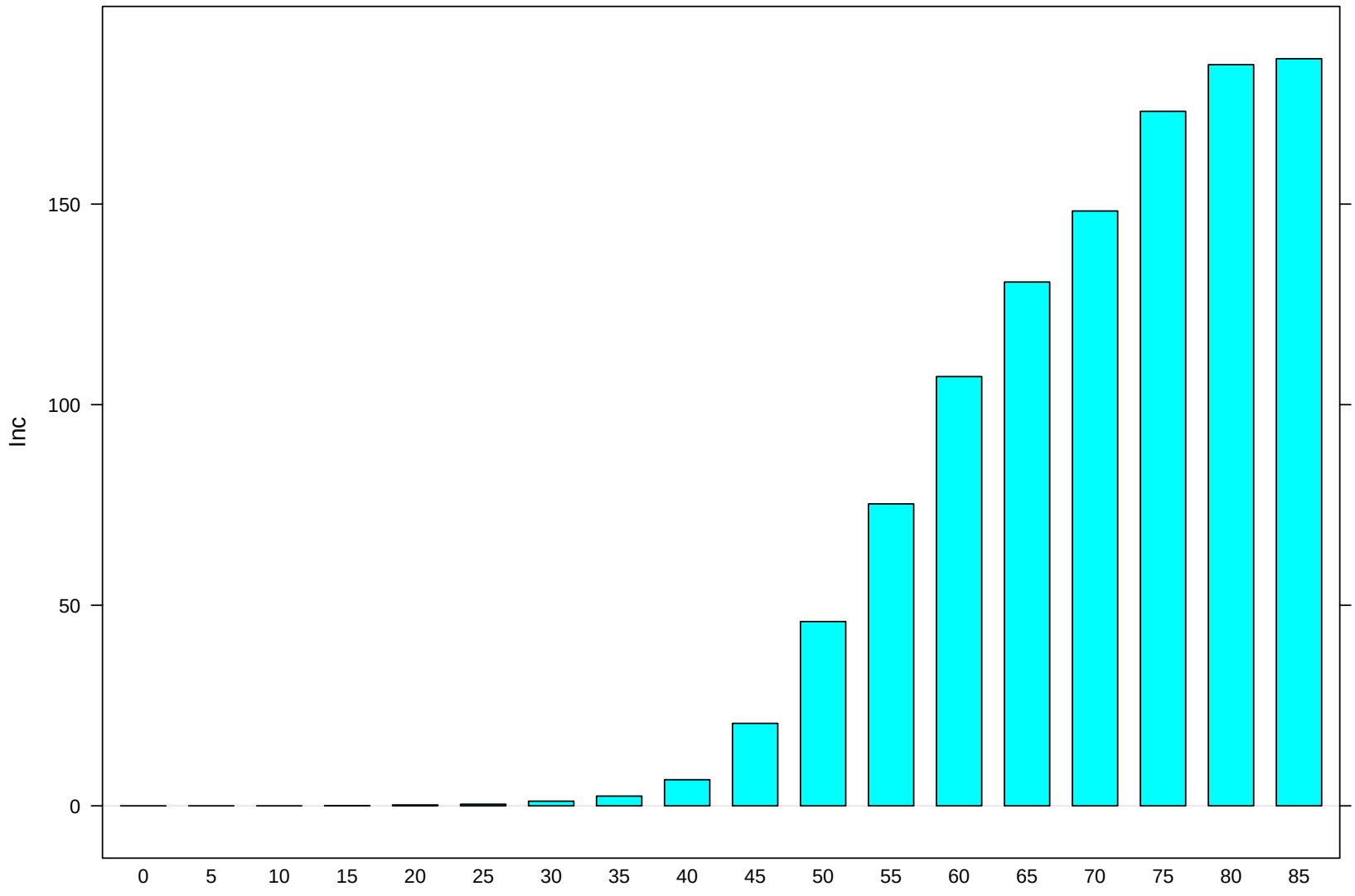
Who are sick?

```
barchart( Inc ~ SEX, data = TimeStratifiedMAJORAMP[ , .( Inc = sum( N )/sum( Population )*100000 ) , .( SEX ) ], origin = 0,
          scales = list( x = list( labels = c( "Férfi", "Nő" ) ) ) )
```



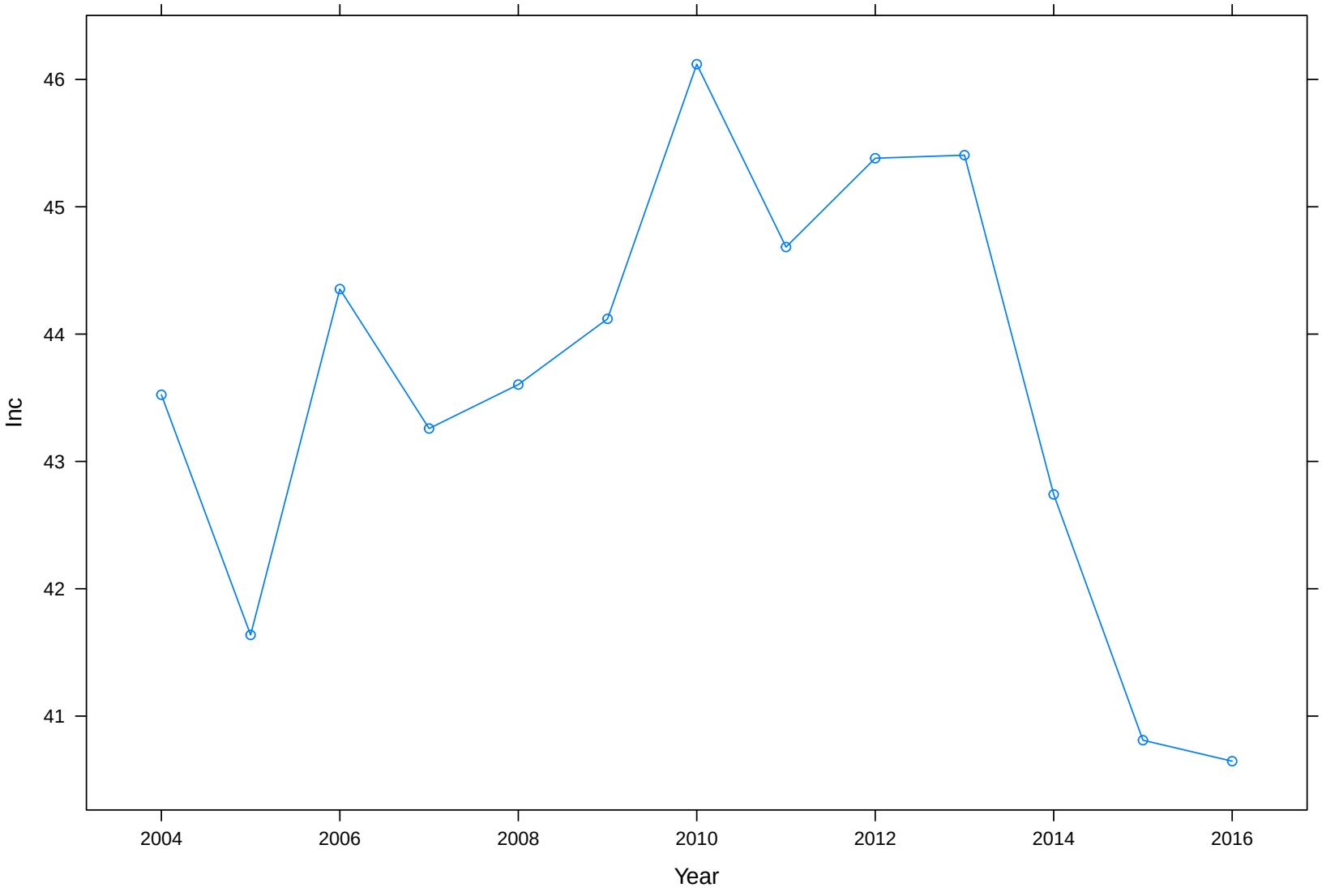
Who are sick?

```
barchart( Inc ~ factor( AgeCut ), data = TimeStratifiedMAJORAMP[ , .( Inc = sum( N )/sum( Population )*100000 ) , .( AgeCut ) ], origin = 0 )
```



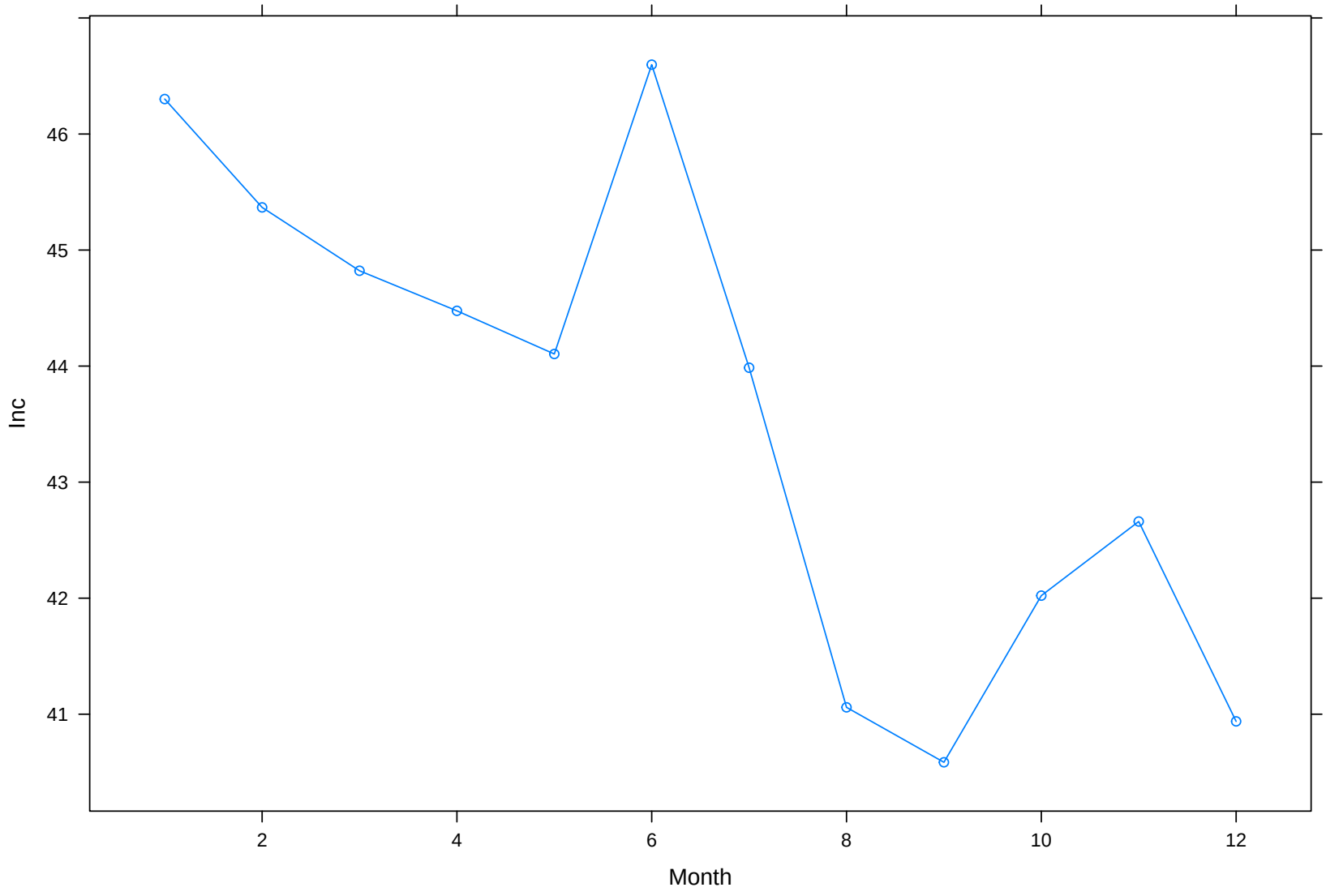
When are people sick?

```
xyplot( Inc ~ Year, data = TimeStratifiedMAJORAMP[ , .( Inc = sum( N )/sum( Population )*100000 ) , .( Year ) ], origin = 0, type = "b" )
```



When are people sick? (another scale)

```
xyplot( Inc ~ Month, data = TimeStratifiedMAJORAMP[ , .( Inc = sum( N )/sum( Population )*100000 ),
                                                    .( Month = as.numeric( format( PROCDATE, "%m" ) ) ) ], origin = 0, type = "b" )
```



Problems of the classical methods

- Hard to take several aspects into account at the same time (hard to estimate, hard to test e.g. for interactions)
- This pertains to the different scales of the same factor
- Control of confounding is only possible through classical methods (standardization)
- Hard to handle continuous variables in an optimal way

Pros and cons of regression models

Pros:

- Easy to take any number of factors into account
- Easy to test their structure
- ''Built-in" standardizaton (again, with testable assumptions)
- Can therefore be used - as every regression model! - to reveal causal relationships from observational data (i.e. to control for confounding)
- Easy to utilize, i.e. to provide meaningful, interpretable results
- Well-known apparatus, wide computational support

Con:

- Model assumptions come in, they have to be tested, whether they are met is a question

Regression models in epidemiology

Typical setup

- We don't have data on those who do not undergo amputation, but we do have data on the size of the population: we will have to stratify
- Outcome (response) is the number of cases (in each strata)
- Predictor variables (that define the strata):
  - Sex, age, etc.
  - Time
  - Location
- Or a combination of these (and in this case possibly including interactions too)

Choice of regression model framework

- Generalized linear model (GLM):
  - Response (Y) follows a distribution from the exponential family
  - A transform (g: link function) of the conditional expected value is modelled linearly:

$$g\left[\mathbb{E}\left(Y|\mathbf{X}\right)\right] = \mathbf{X}^T\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

  - If needed, the variance should also be specified
- Response is count data
  - Poisson (link: log, variance not needed)
  - Negative binomial, typical parametrization is such that variance for μ expected value is $\mu + \mu^2/\theta$ (link: log, θ needs to be estimated, possibly to do through ML or REML)

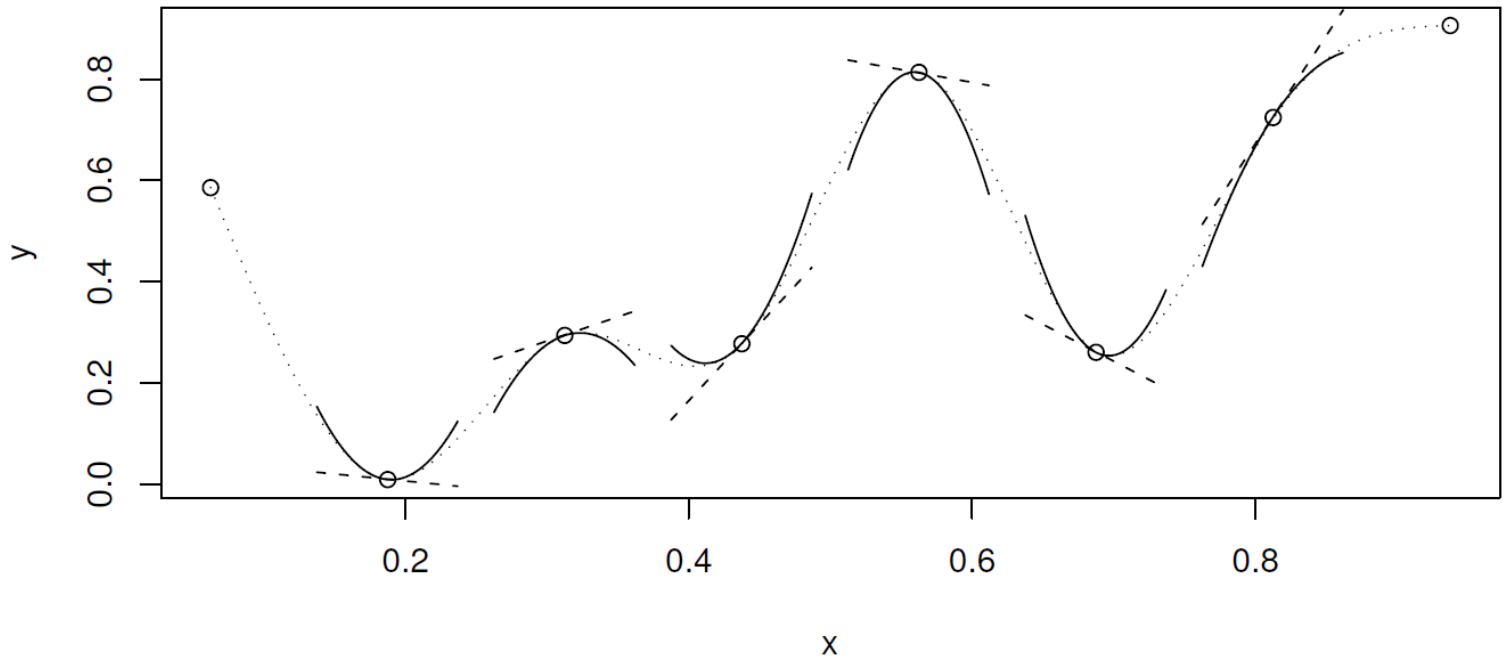
Choice of functional form

- Parametric: $\beta_0 + \beta_{Year}\, Year$
- Non-parametric: $\beta_0 + \beta_{Year=2005} D_{Year=2005} + \beta_{Year=2006} D_{Year=2006} + \ldots + \beta_{Year=2014} D_{Year=2014}$
- ''Semi-parametric'': splines

Questions: flexibility (forcing external assumptions on the data), extrapolation, interpretation, efficiency when estimating

Sidenote: splines

- Technically parametric, but extremely flexible, yet consumes little df
- Definition in one sentence: piecewise defined polynomial (with some constraints so that they meet in knots, and also the transition is smooth)
- Precisely: when order p polynomials are used, we require derivates up to order $p - 1$ to be the same in the knots (plus some requirement at the endpoints)

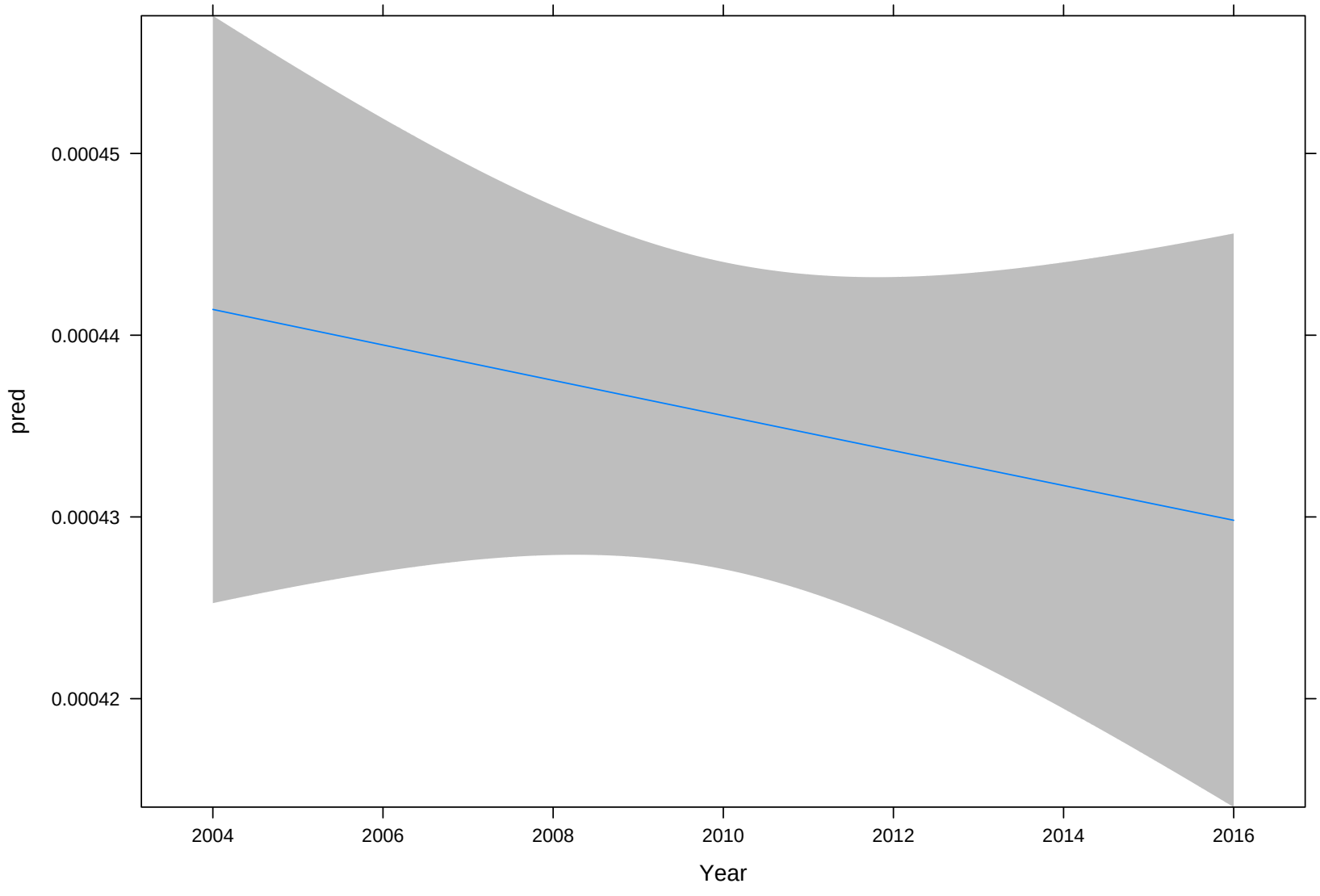


Sidenote: splines

- Typically $p = 3$ and linear at the endpoint: restricted cubic spline
- We have to select number and location of the knots...
- ... and then estimate $k - 1$ parameters (for k knots)
- Can be reduced to simple regression
- Number of knots is usually set to ''high enough", overfitting is avoided by penalized estimation (which penalizes too wiggly functional form)
- With this, precise knot location doesn't really matter either
- Wiggliness: integral of the square of the second derivative
- Penalized estimation: how well the curve fits plus λ times wiggliness
- This can also be reduced to simple regression!
- λ is usually selected with cross validation

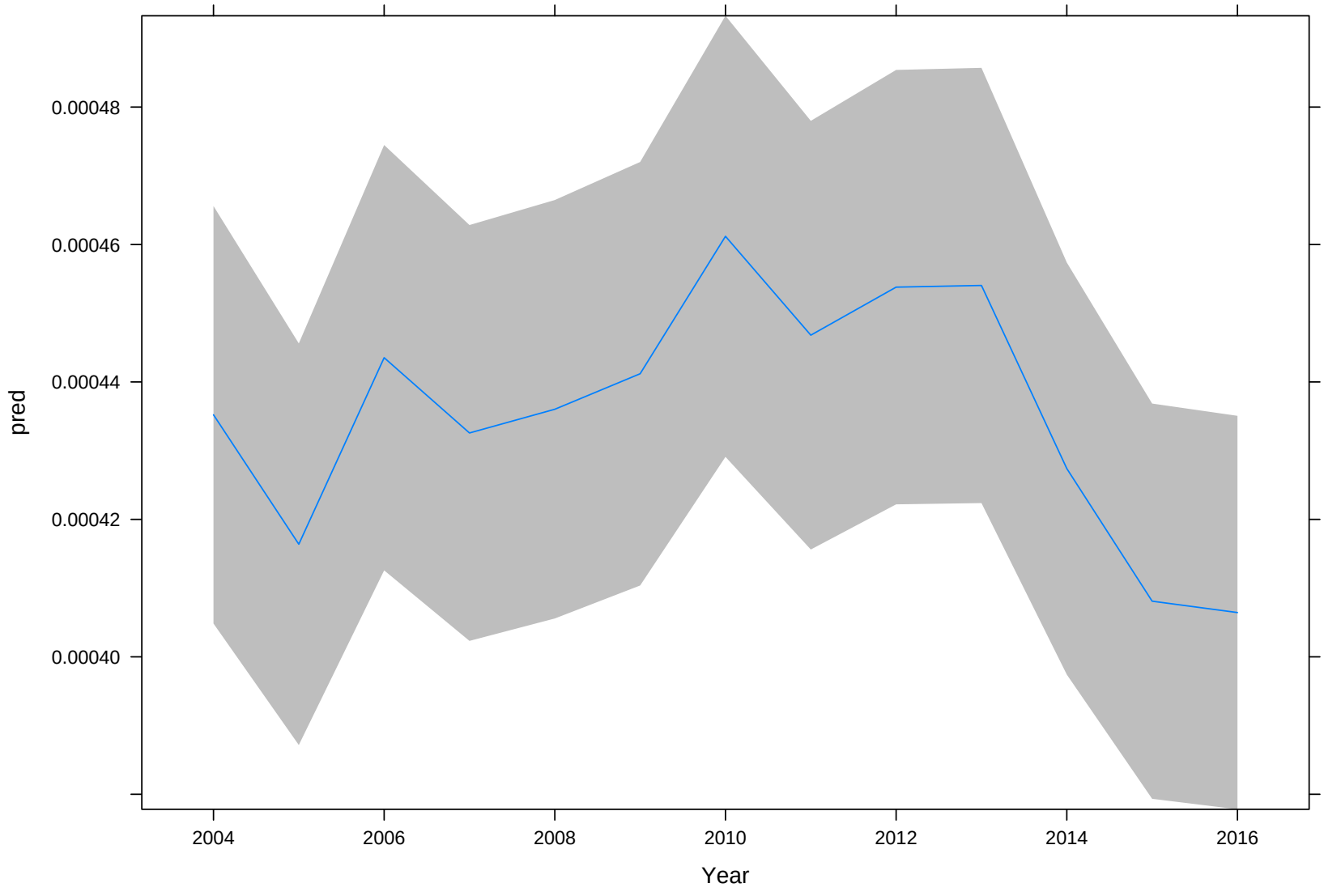
Choice of functional form: parametric

```
fit <- gam( N ~ Year, offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
pred <- data.frame( predict( fit, data.frame( Year = yrgrid ), se.fit = TRUE, type = "response" ) )
pred <- with( pred, data.frame( cilwr = fit-qnorm( 0.975 )*se.fit, pred = fit, ciupr = fit+qnorm( 0.975 )*se.fit, Year = yrgrid ) )
xYplot( Cbind( pred, cilwr, ciupr ) ~ Year, data = pred, method = "filled bands", type = "l", col.fill = "gray" )
```



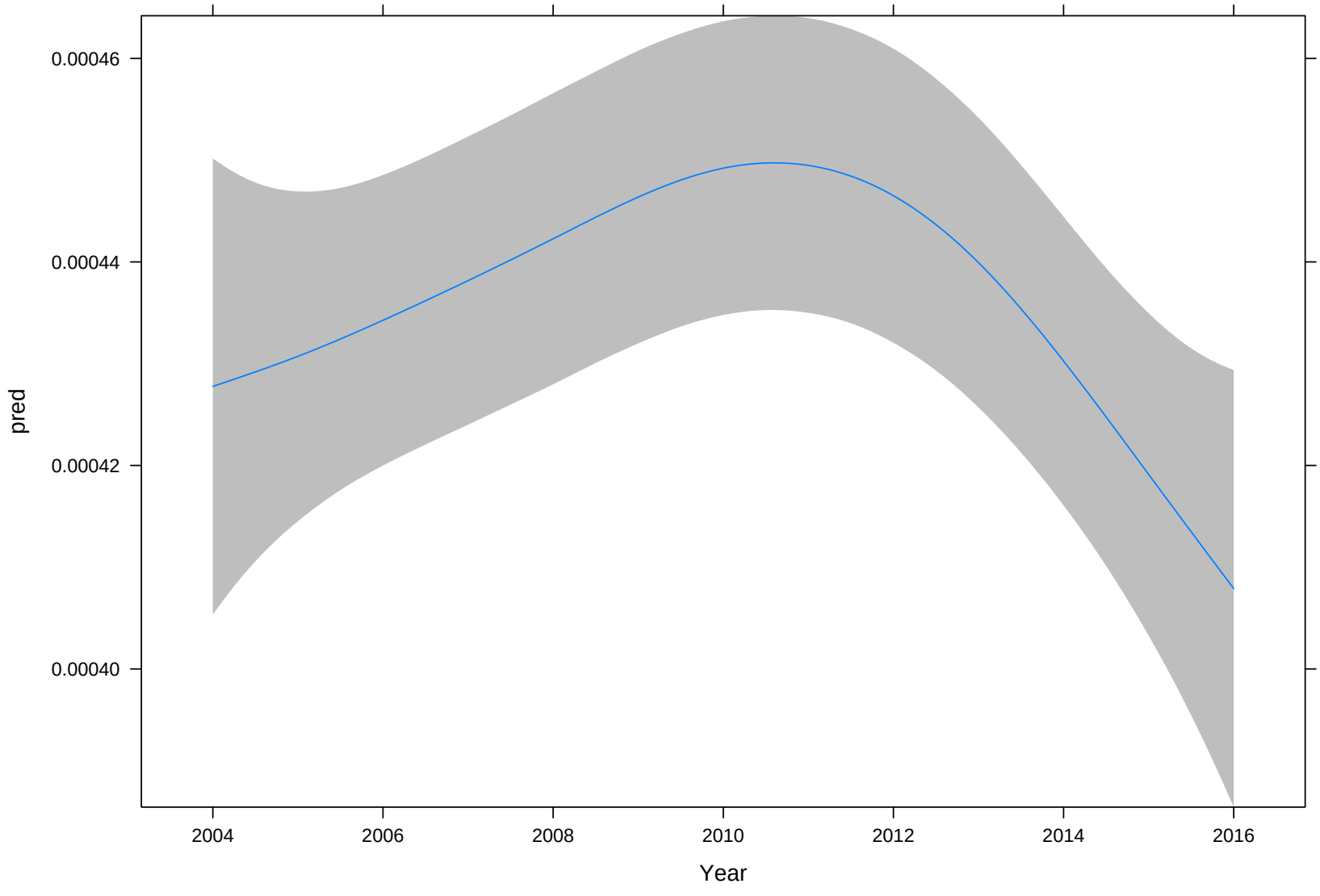
Choice of functional form: non-parametric

```
fit <- gam( N ~ as.factor( Year ), offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
pred <- data.frame( predict( fit, data.frame( Year = 2004:2016 ), se.fit = TRUE, type = "response" ) )
pred <- with( pred, data.frame( cilwr = fit-qnorm( 0.975 )*se.fit, pred = fit, ciupr = fit+qnorm( 0.975 )*se.fit, Year = 2004:2016 ) )
xYplot( Cbind( pred, cilwr, ciupr ) ~ Year, data = pred, method = "filled bands", type = "l", col.fill = "gray" )
```



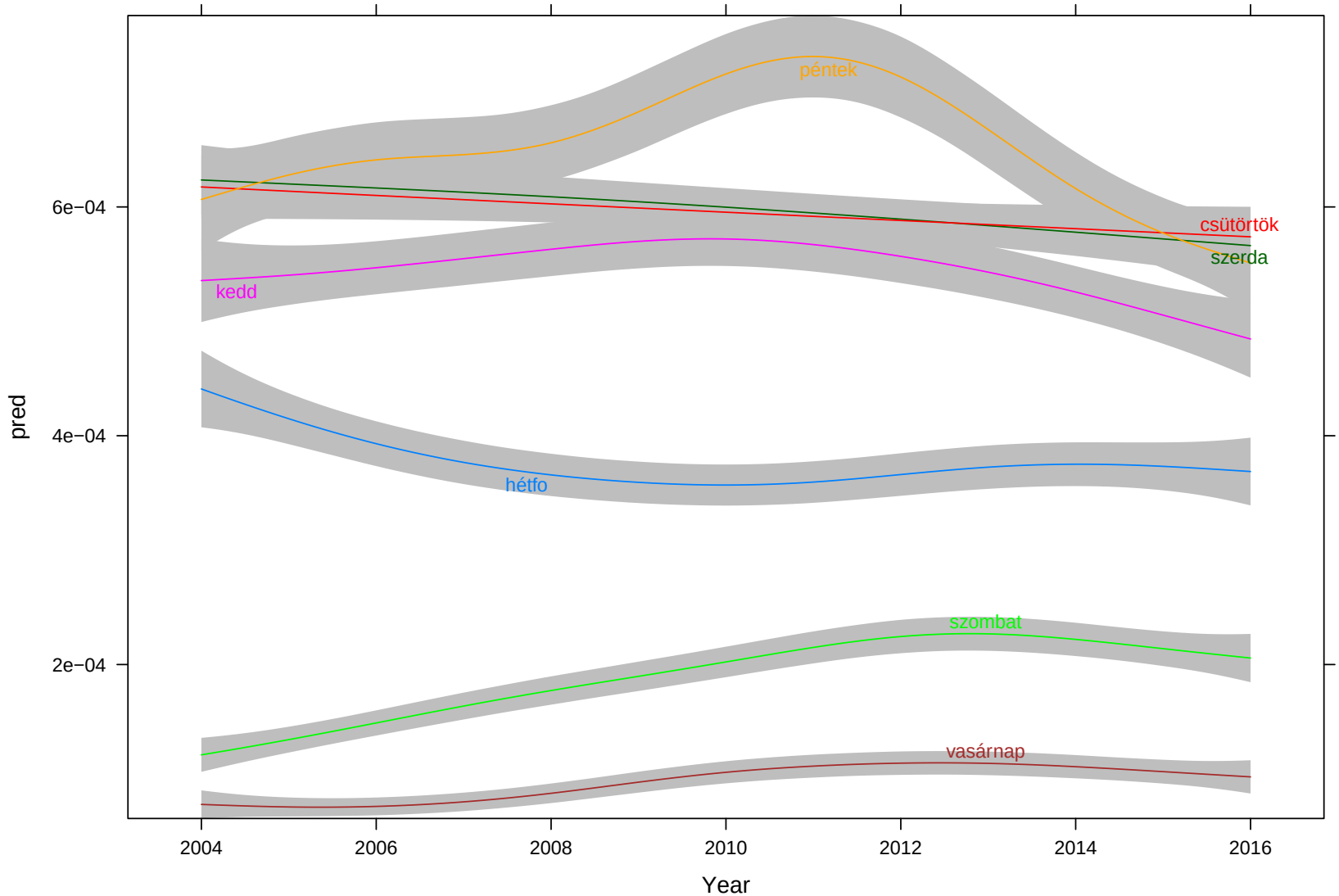
Choice of functional form: semi-parametric

```
fit <- gam( N ~ s( Year ), offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
pred <- data.frame( predict( fit, data.frame( Year = yrgrid ), se.fit = TRUE, type = "response" ) )
pred <- with( pred, data.frame( cilwr = fit-qnorm( 0.975 )*se.fit, pred = fit, ciupr = fit+qnorm( 0.975 )*se.fit, Year = yrgrid ) )
xYplot( Cbind( pred, cilwr, ciupr ) ~ Year, data = pred, method = "filled bands", type = "l", col.fill = "gray" )
```



A model with interaction between year and day-of-week

```
fit <- gam( N ~ s( Year, by = DOW ) + DOW, offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
predgrid <- expand.grid( DOW = unique( TimeStratified2$DOW ), Year = yrgrid )
pred <- data.frame( predict( fit, predgrid, se.fit = TRUE, type = "response" ) )
pred <- with( pred, cbind( data.frame( cilwr = fit-qnorm( 0.975 )*se.fit, pred = fit, ciupr = fit+qnorm( 0.975 )*se.fit ), predgrid ) )
xYplot( Cbind( pred, cilwr, ciupr ) ~ Year, groups = DOW, data = pred[ order( pred$Year ), ], method = "filled bands", type = "l", col.fill = "gray" )
```



Creating epidemiologic models

Iterative process:

- Model specification (how the model looks like - we just covered a few important aspects of it)
- Model diagnostics (which model is acceptable?)
- Model selection (which model is the best?)

Model diagnostics and -specification: the detective work

Model diagnostics

- Checking whether model assumptions are met
- Important ones in our current issue:
  - Adequacy of the functional form
  - Non-autocorrelation of the residuals
  - Possible overdispersion

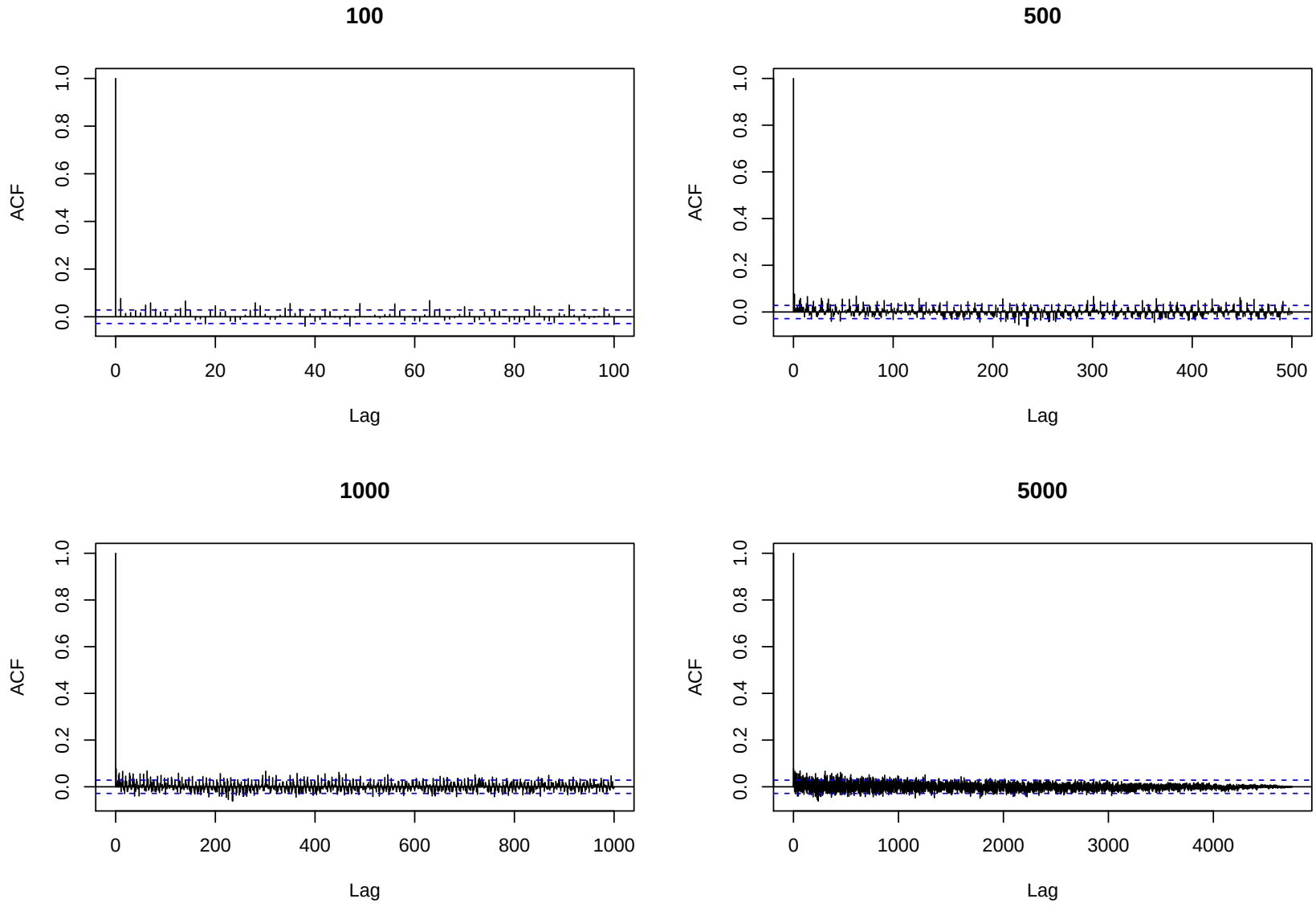
Initial model

- Collapsed across sex- and age-strata
- Explanatory variables:
  - Day-of-week (DOW)
  - Non-working day (NWD)
  - Swapped working day
  - Surgery congress
  - Time (continuously, not yearly!) to model secular trends
  - Day-of-year (seasonality)

```
fit <- gam( N ~ DOW + Swap + NWD + SurgeryCongress + s( PROCDATEnum ) + s( DOY, bs = "cc" ),
            offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
```

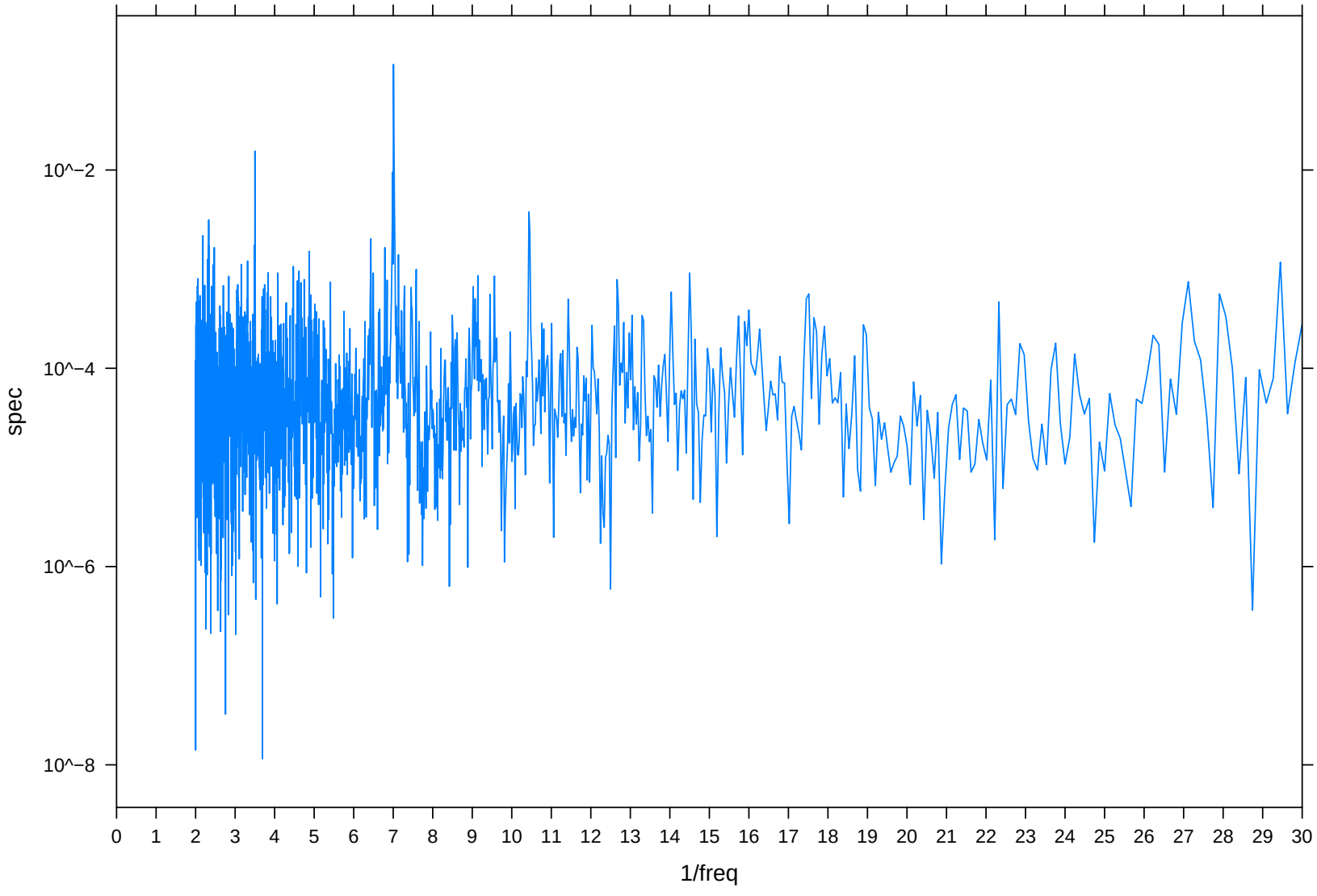
Autocorrelation of the inital model

```
par( mfrow = c( 2, 2 ) )
for( l in c( 100, 500, 1000, 5000 ) ) acf( resid( fit ), lag.max = l, main = l )
```



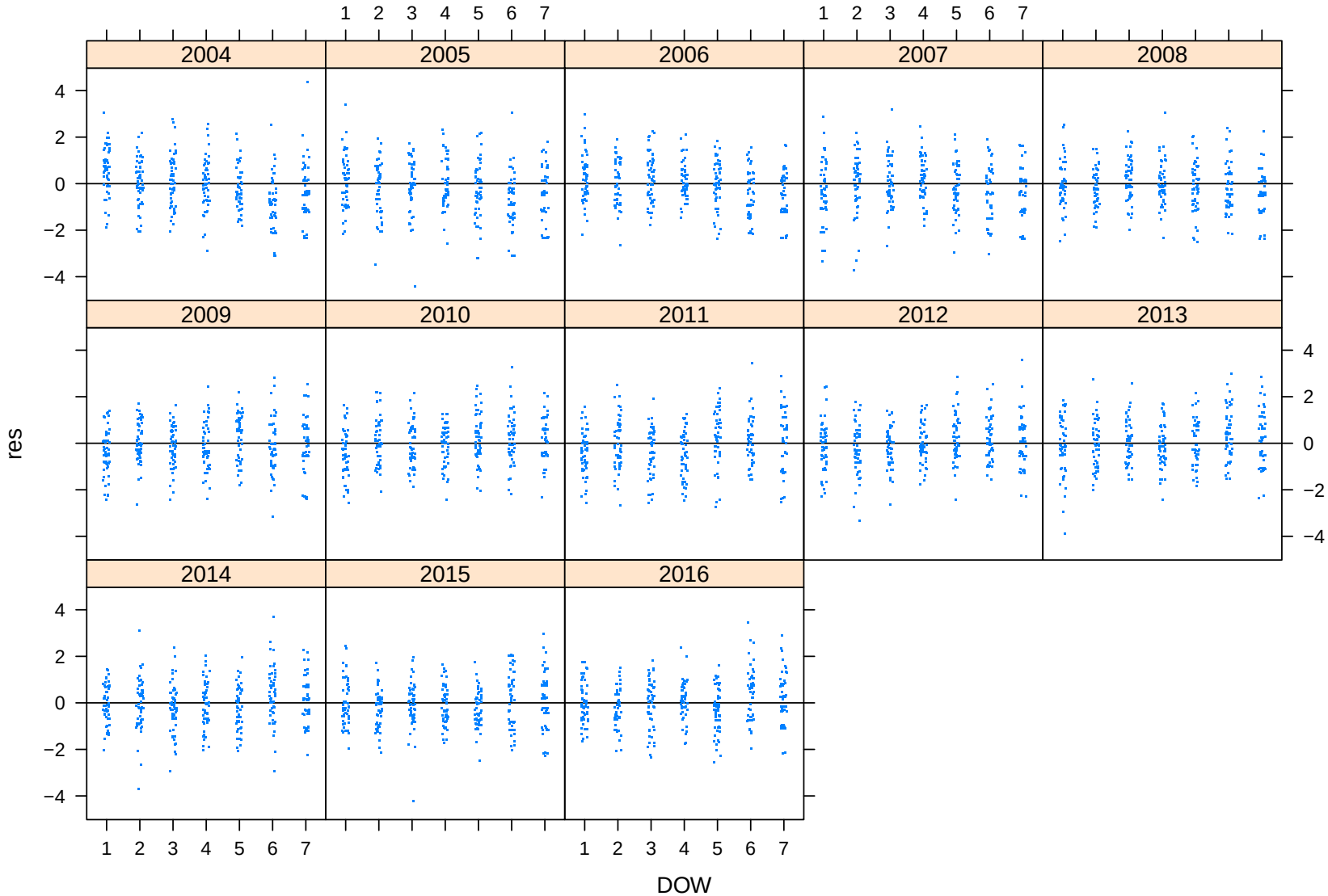
Spectrum of the ACF

```
xyplot( spec~1/freq, data = spectrum( acf( resid( fit ), lag.max = 5000,
                                          plot = FALSE )$acf, plot = FALSE ),
        scale = list( y = list( log = 10 ), x = list( at = 0:30 ) ), xlim = c( 0, 30 ),
        type = "l" )
```



Residuals per DOW

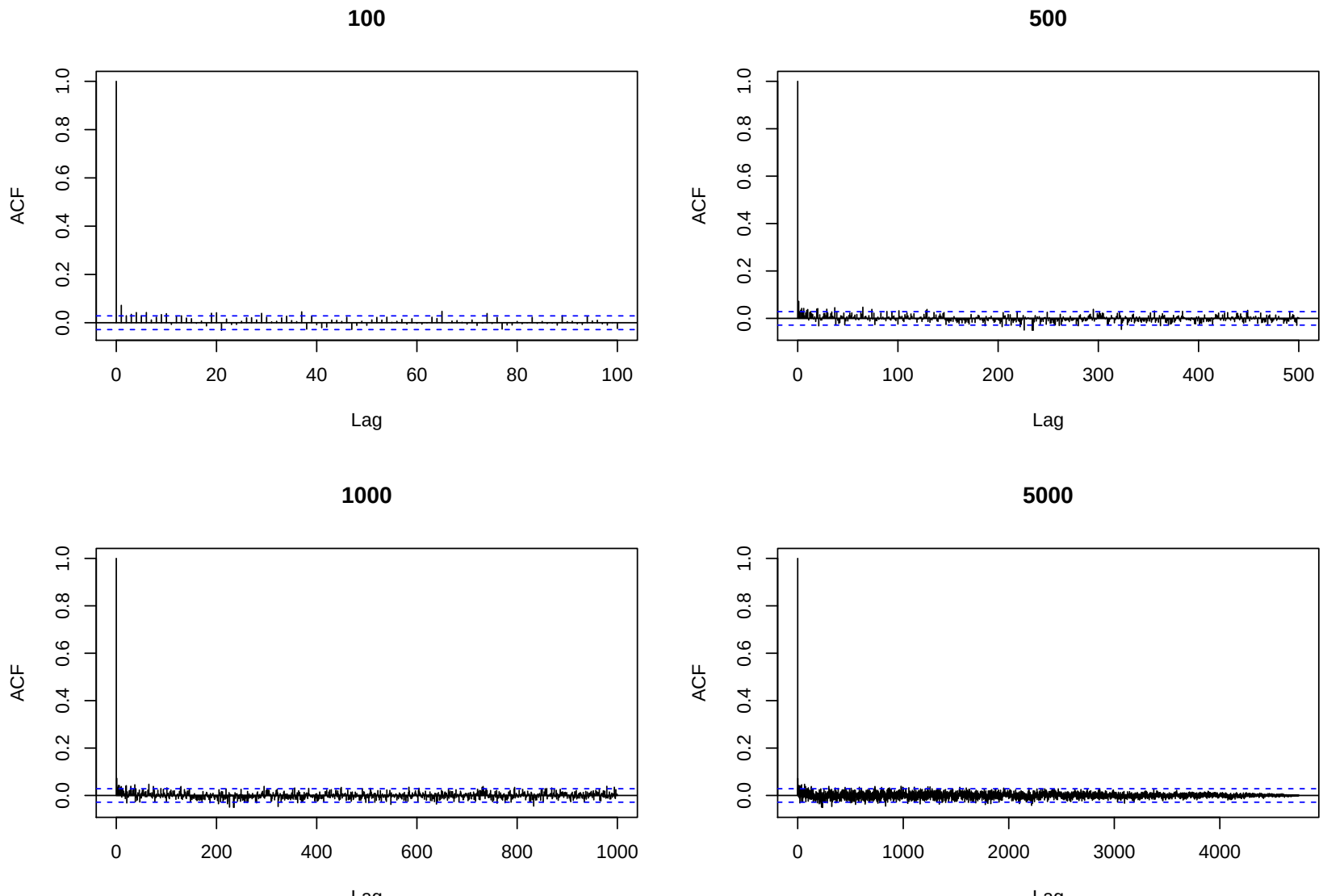
```
xyplot( res ~ DOW | as.factor( Year ),
        data = cbind( TimeStratified2, res = resid( fit ) ), jitter.x = TRUE,
        scales = list( x = list( labels = 1:7 ) ), as.table = TRUE, pch = ".",
        abline = list( h = 0 ) )
```



Interaction between DOW and long-term trend

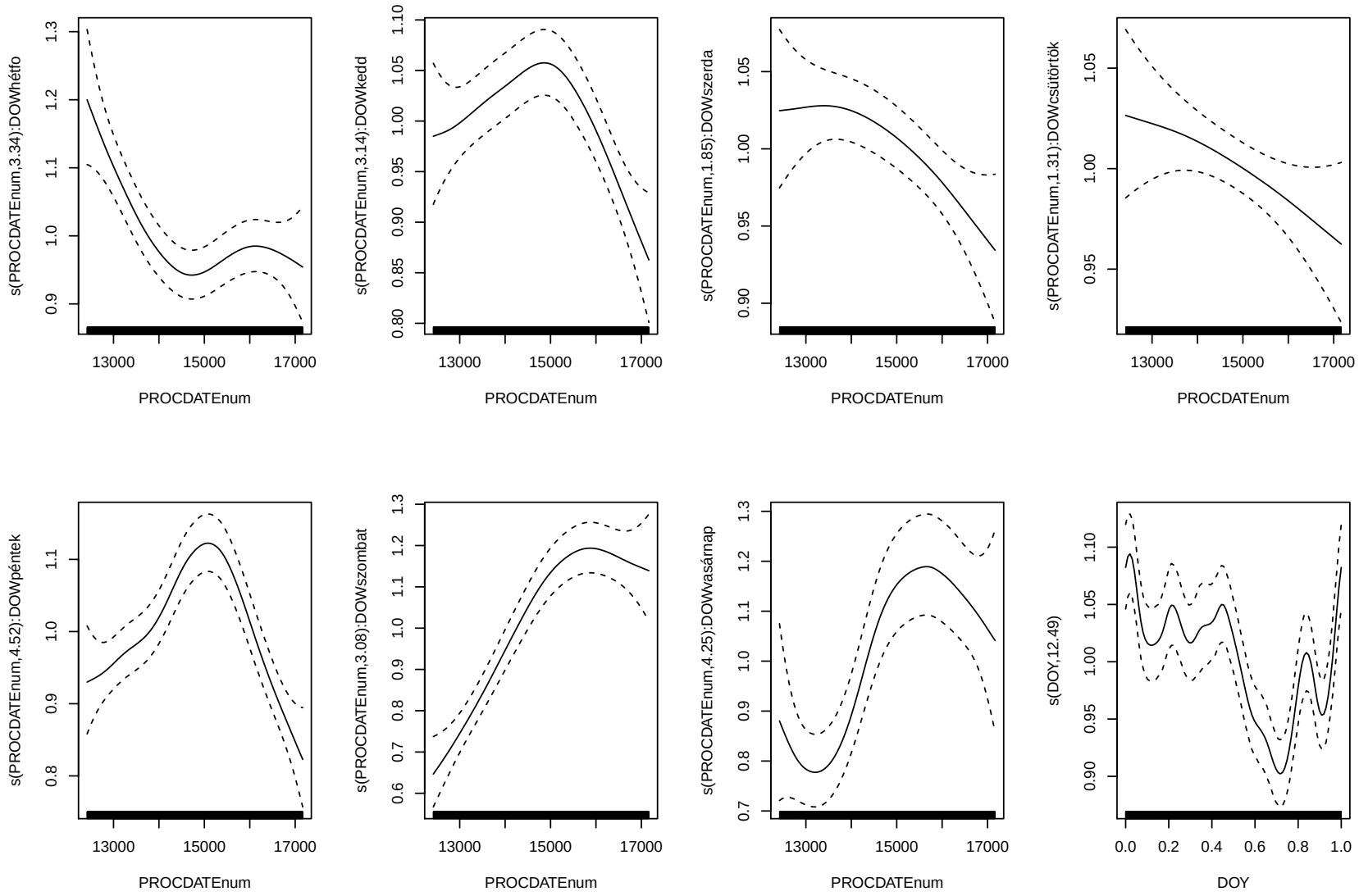
```
fit <- gam( N ~ DOW + Swap + NWD + CataractaCongress + SurgeryCongress +
              CardiologyCongress + s( PROCDATEnum, by = DOW ) + s( DOY, bs = "cc" ),
            offset = log( Population ), data = TimeStratified2,
            family = nb( link = log ) )
```

```
par( mfrow = c( 2, 2 ) )
for( l in c( 100, 500, 1000, 5000 ) ) acf( resid( fit ), lag.max = l, main = l )
```



Checking the adequacy of the basis dimension and having a look at the model

```
fit <- gam( N ~ DOW + Swap + NWD + CataractaCongress + SurgeryCongress + CardiologyCongress + s( PROCDATEnum, by = DOW, k = 30 ) +
              s( DOY, bs = "cc", k = 30 ), offset = log( Population ), data = TimeStratified2, family = nb( link = log ) )
```



Most outlying residuals

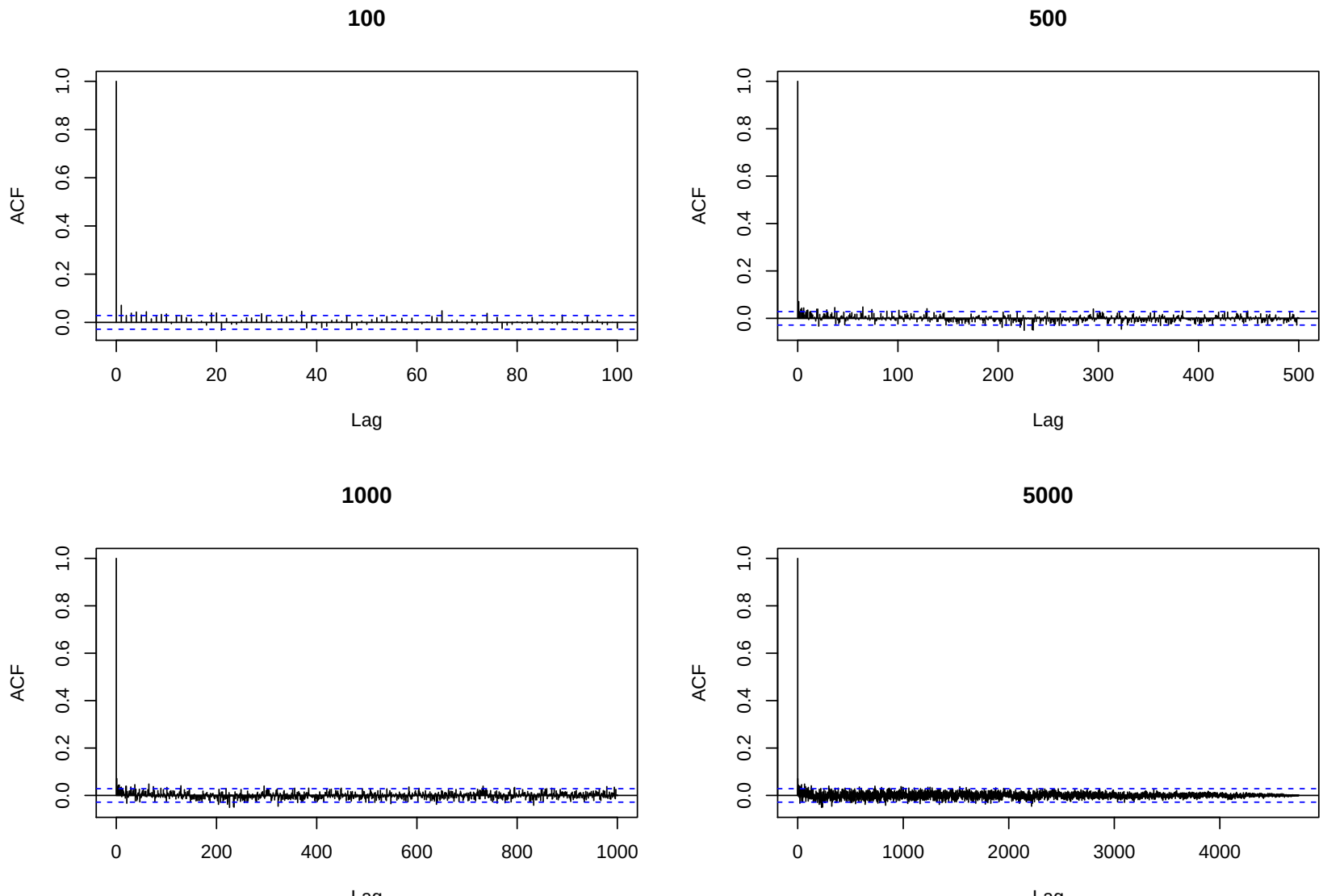
```
res <- data.frame( TimeStratified2, res = resid( fit ) )
res <- res[ order( abs( res$res ), decreasing = TRUE ), c( "PROCDATE", "res", "N" ) ]
head( res, 10 )
```

	PROCDATE	res	N
81	2004-03-21	4.663787	13
455	2005-03-30	-4.344820	2
4200	2015-07-01	-4.316001	2
675	2005-11-05	4.100290	26
3470	2013-07-01	-3.835463	1
1217	2007-05-01	-3.810736	0
3835	2014-07-01	-3.676288	3
454	2005-03-29	-3.562019	1
3958	2014-11-01	3.415572	9
3772	2014-04-29	3.394891	31

Handling Semmelweis Day

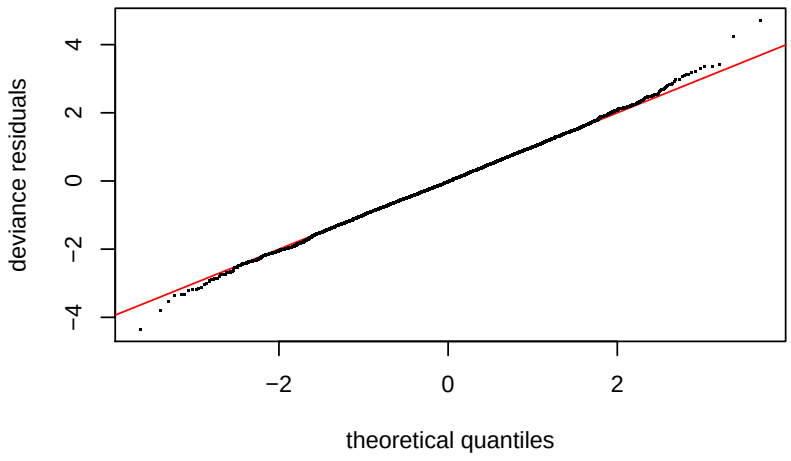
```
fit <- gam( N ~ DOW + Swap + NWD + CataractaCongress + SurgeryCongress + SemmelweisDay +
              CardiologyCongress + s( PROCDATEnum, by = DOW ) + s( DOY, bs = "cc" ),
            offset = log( Population ), data = TimeStratified2,
            family = nb( link = log ) )
```

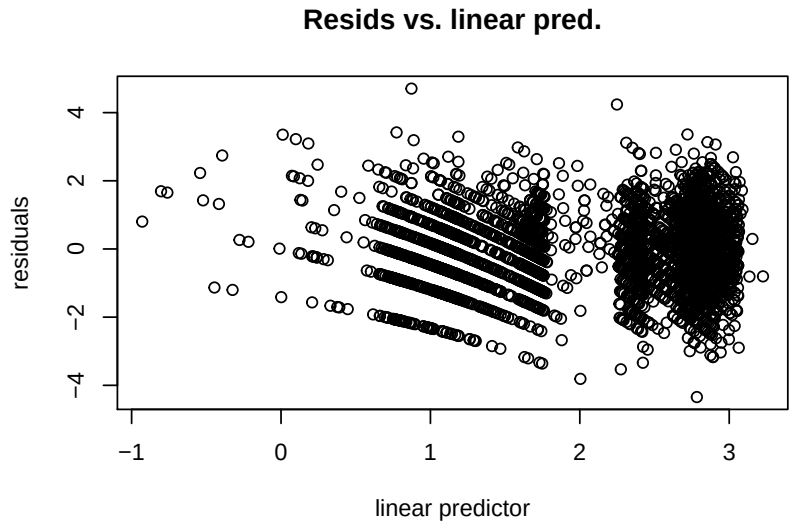
```
par( mfrow = c( 2, 2 ) )
for( l in c( 100, 500, 1000, 5000 ) ) acf( resid( fit ), lag.max = l, main = l )
```

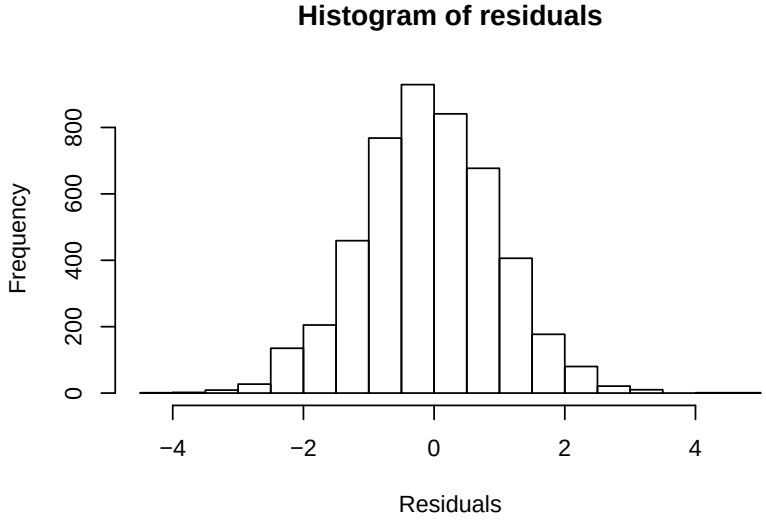


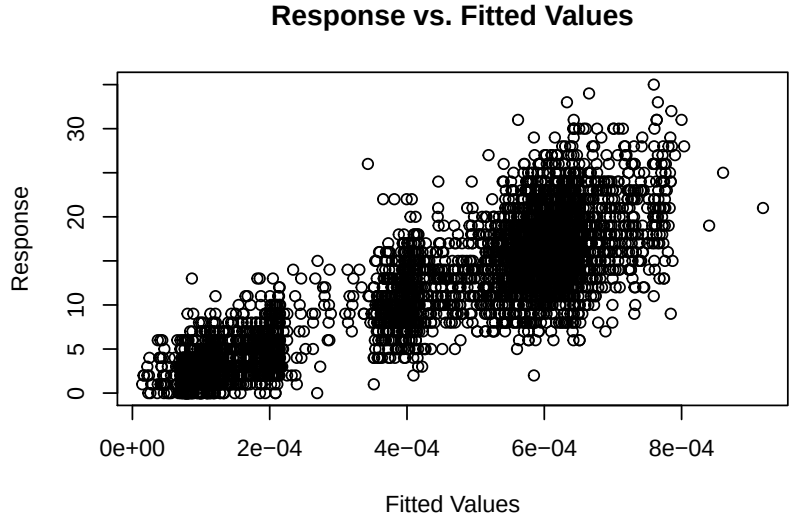
Dispersion of the residuals

```
par( mfrow = c( 2, 2 ) )
gam.check( fit )
```









Including age and sex

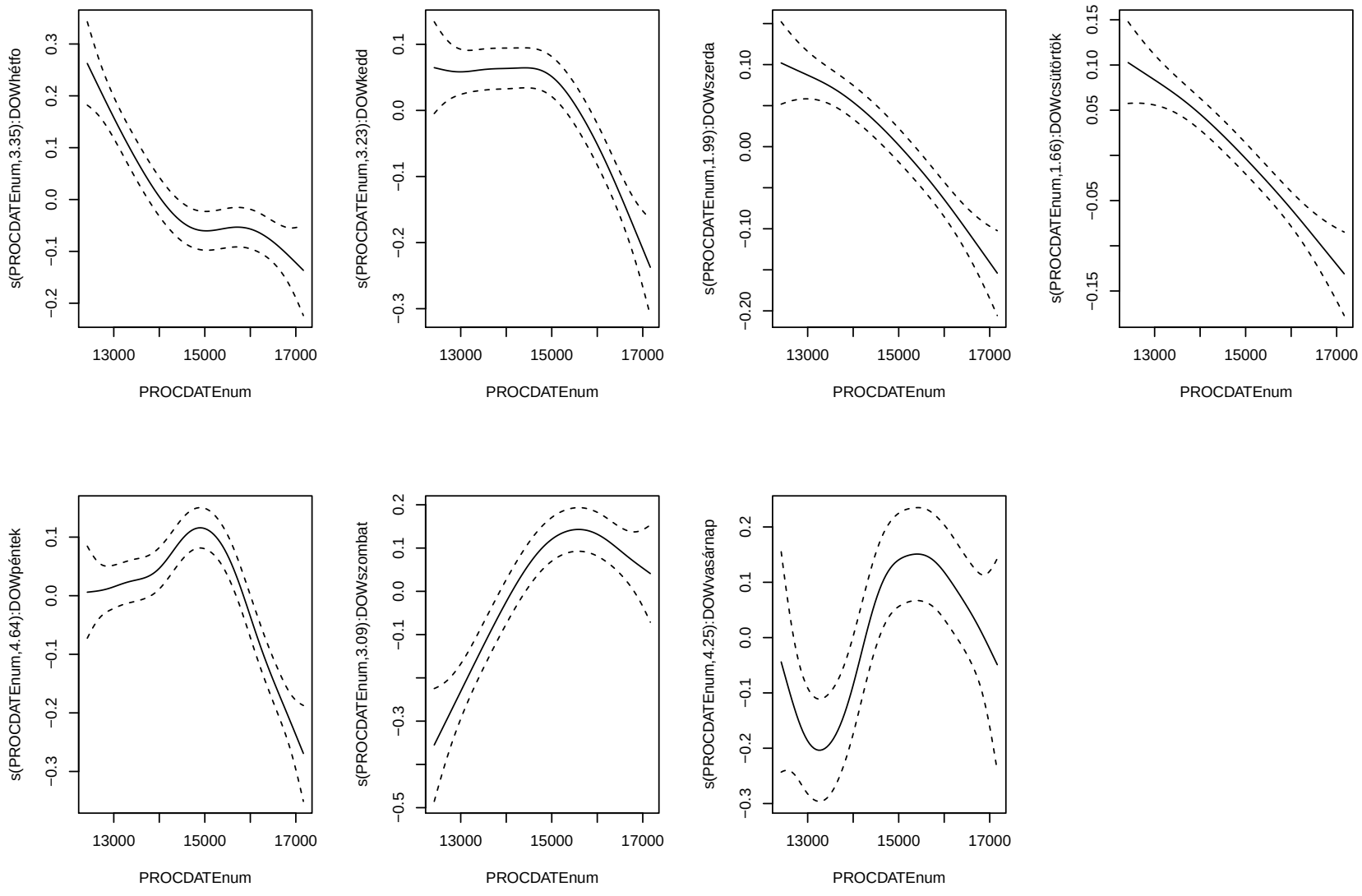
- Problem with all these analyis is that it omits age and sex...
- But the compisition of the population according to these can change in decade long intervals
- Classical solution: epidemiologic standardization
- We will now use a modern alternative: including these as covariates in the regression!
- Of course we will need a more detailed age pyramid

```
fit <- gam( N ~ DOW + Swap + NWD + SurgeryCongress + SemmelweisDay + s( PROCDATEnum, by = DOW ) +
              s( DOY, bs = "cc" ) + s( AgeCut, by = SEX ) + SEX,
            offset = log( Population ), data = TimeStratifiedMAJORAMP,
            family = nb( link = log ) )
```

Results: using the obtained model

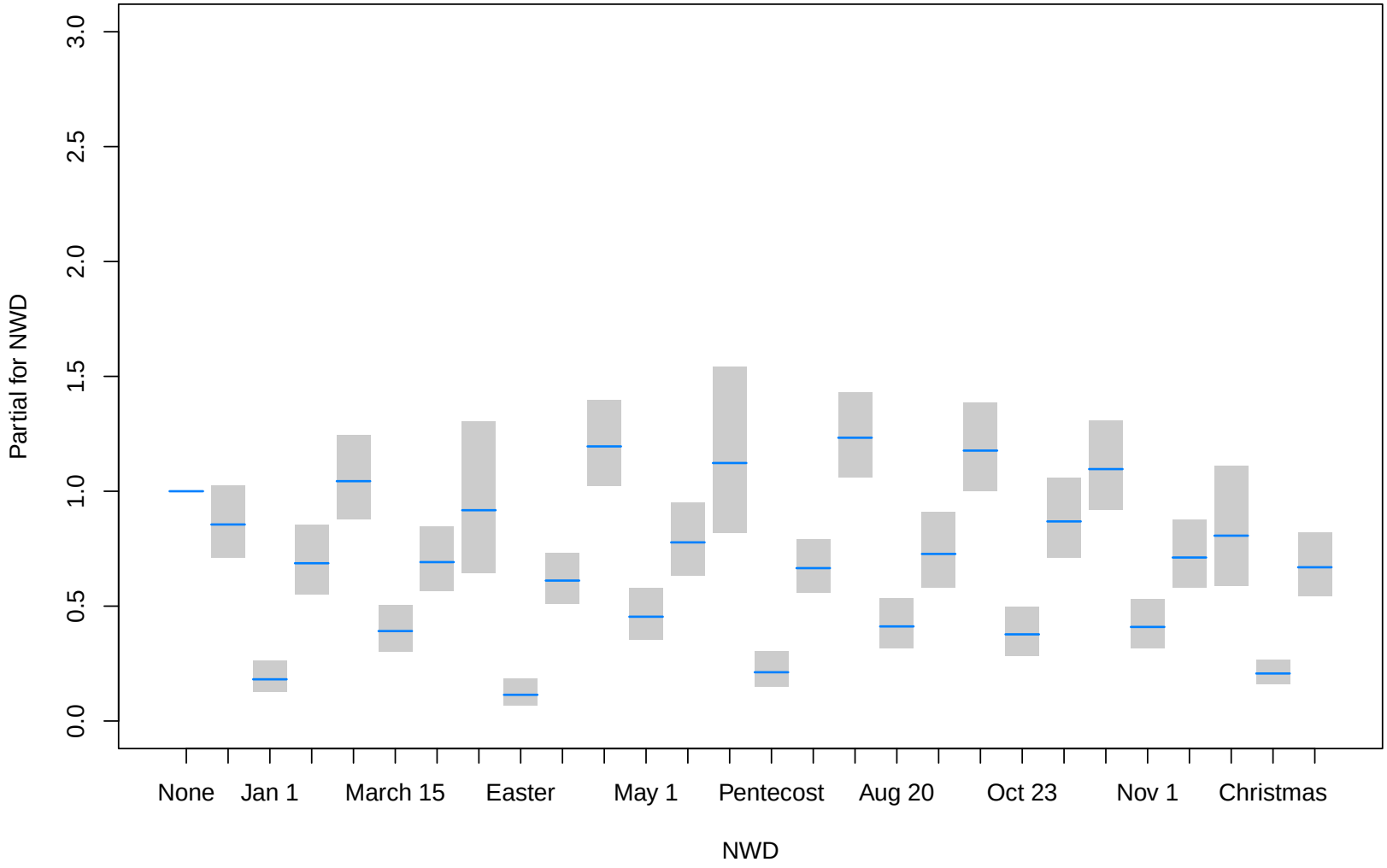
Long-term trend and DOW

```
par( mfrow = c( 2, 4 ) )
for( i in 1:7 )
  plot( fit, select = i, scale = 0 )
```



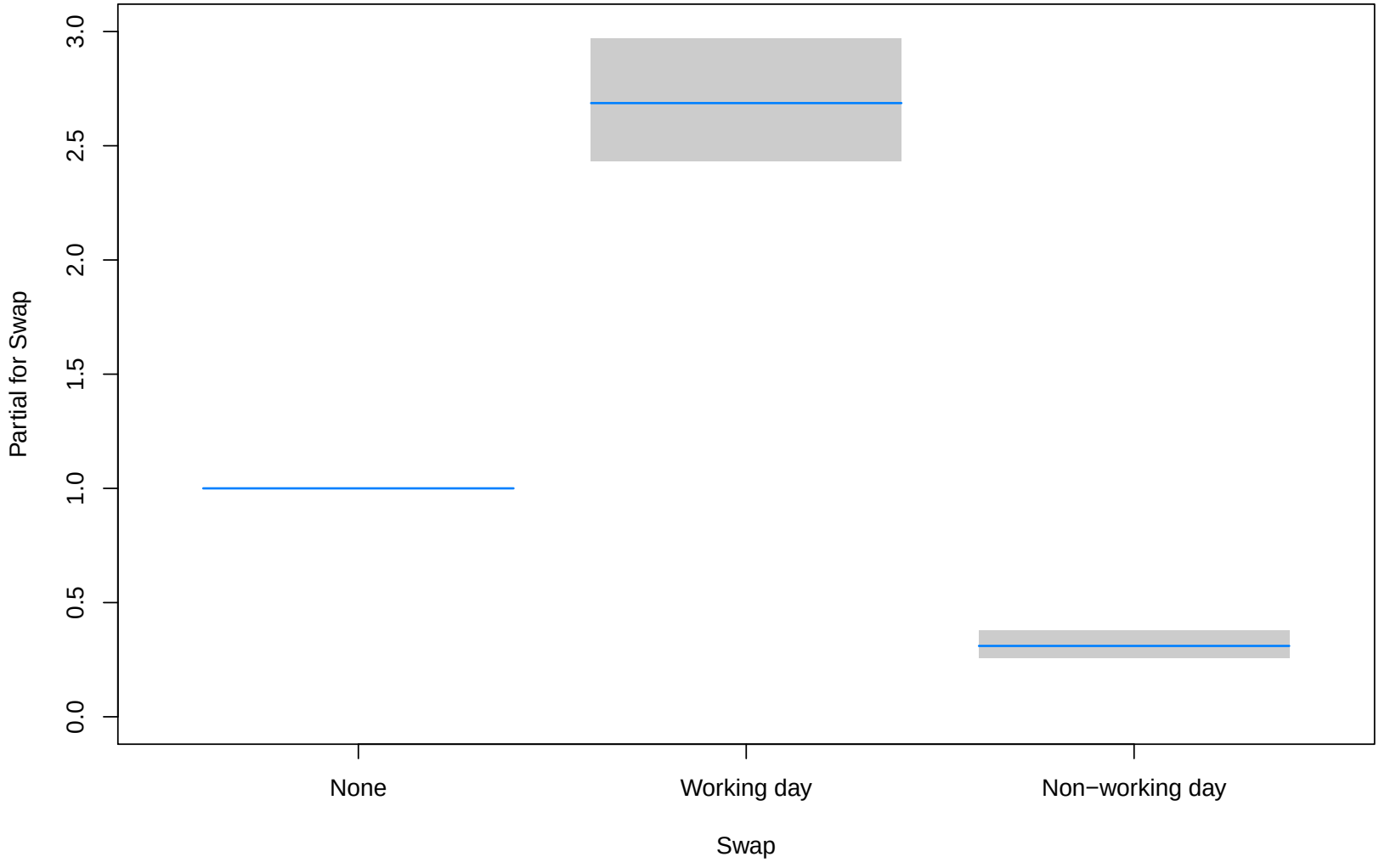
Non-working day

```
termplot2( fit, terms = "NWD", se = TRUE, yscale = "exponential", col.term = trellis.par.get()$superpose.line$col[1],
           col.se = "gray80", se.type = "polygon", ylim = c( -Inf, log( 3 ) ) )
```



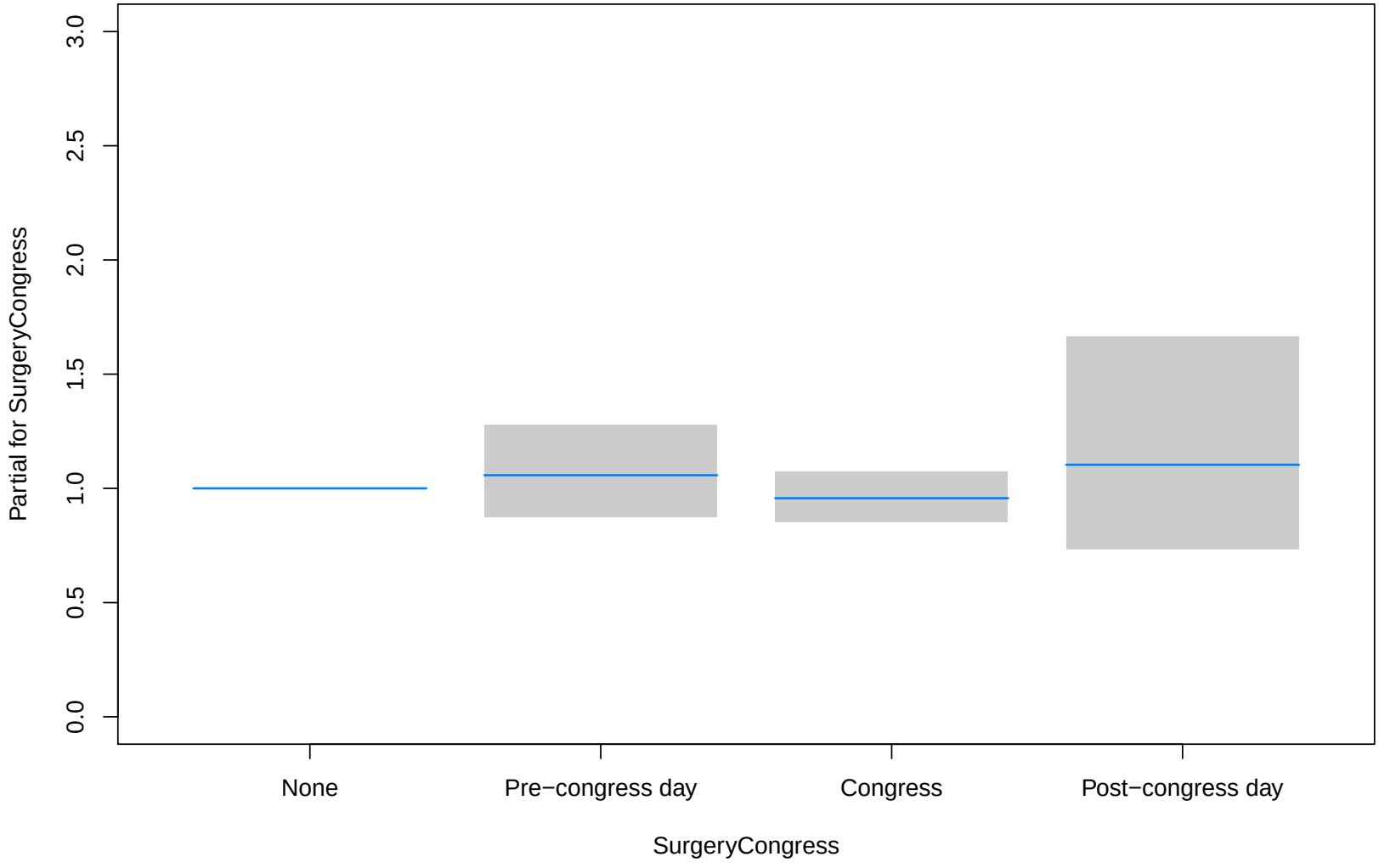
Swapped working day

```
termplot2( fit, terms = "Swap", se = TRUE, yscale = "exponential", col.term = trellis.par.get()$superpose.line$col[1],
           col.se = "gray80", se.type = "polygon", ylim = c( -Inf, log( 3 ) ) )
```



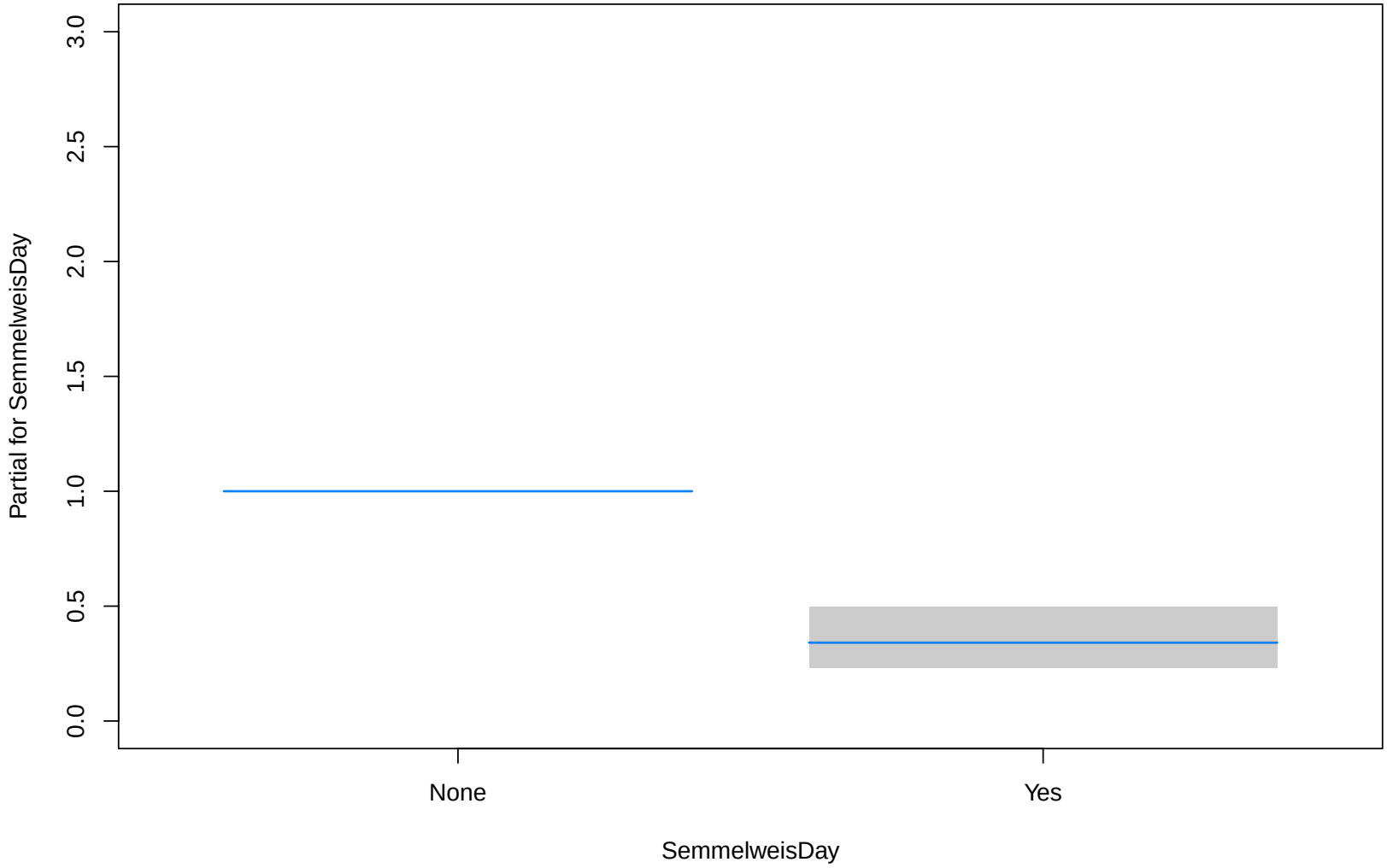
Surgery congress

```
termplot2( fit, terms = "SurgeryCongress", se = TRUE, yscale = "exponential", col.term = trellis.par.get()$superpose.line$col[1],
           col.se = "gray80", se.type = "polygon", ylim = c( -Inf, log( 3 ) ) )
```



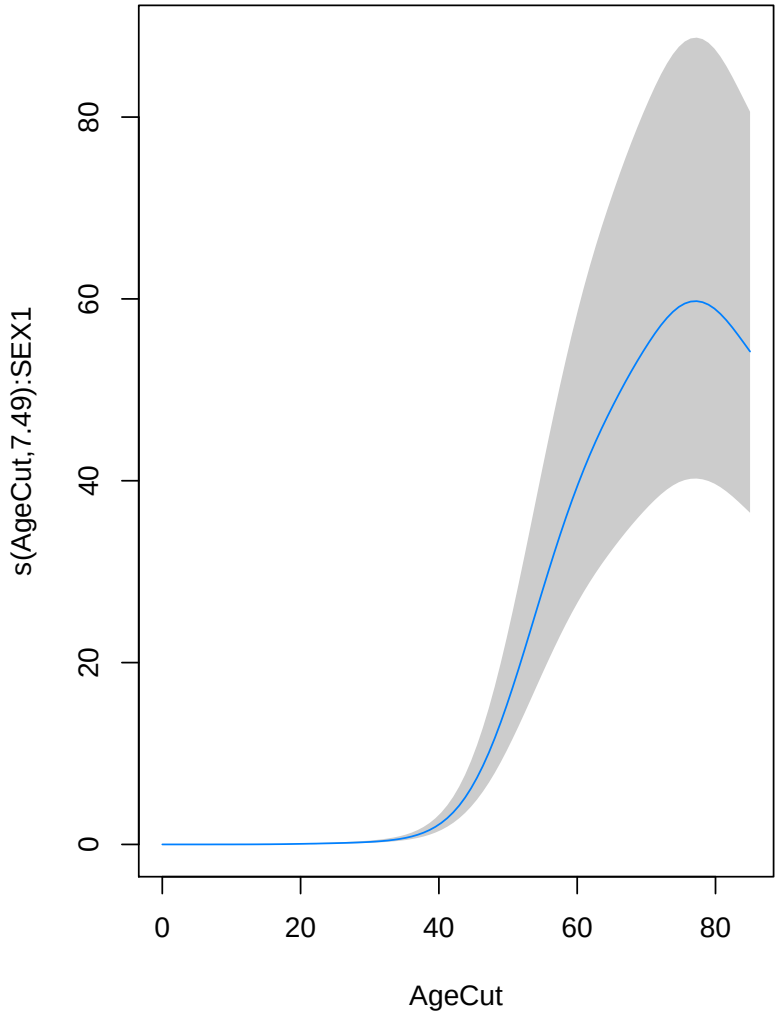
Semmelweis Day

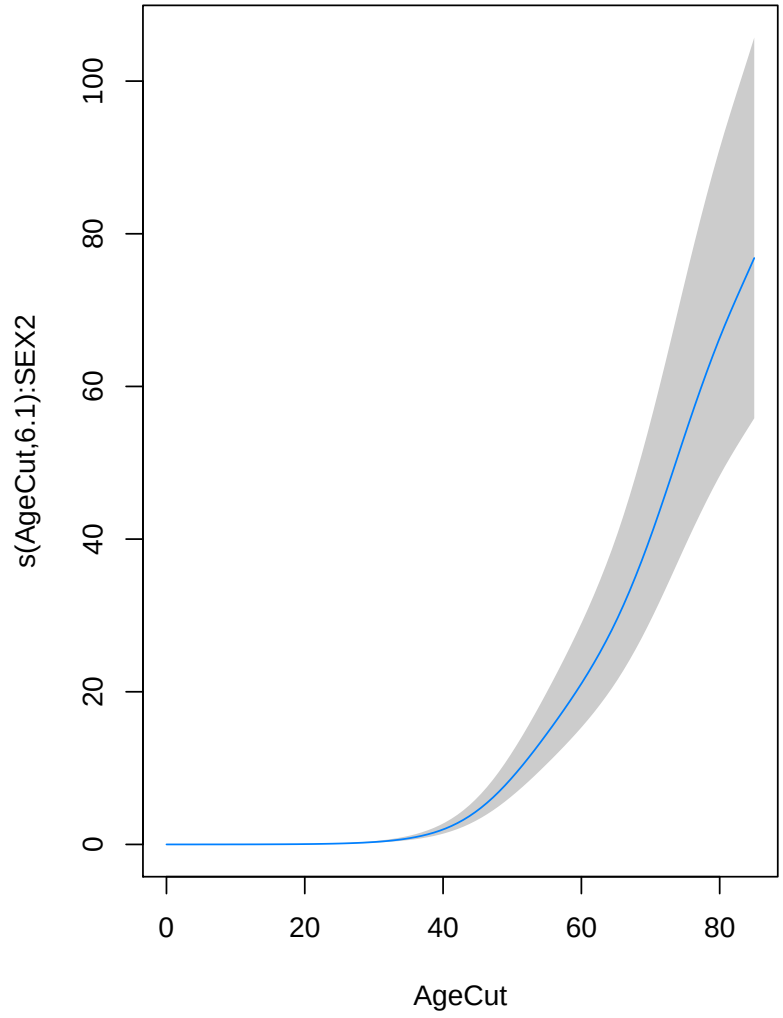
```
termplot2( fit, terms = "SemmelweisDay", se = TRUE, yscale = "exponential", col.term = trellis.par.get()$superpose.line$col[1],
           col.se = "gray80", se.type = "polygon", ylim = c( -Inf, log( 3 ) ) )
```



Effect of age and sex (estimated in integrated manner)

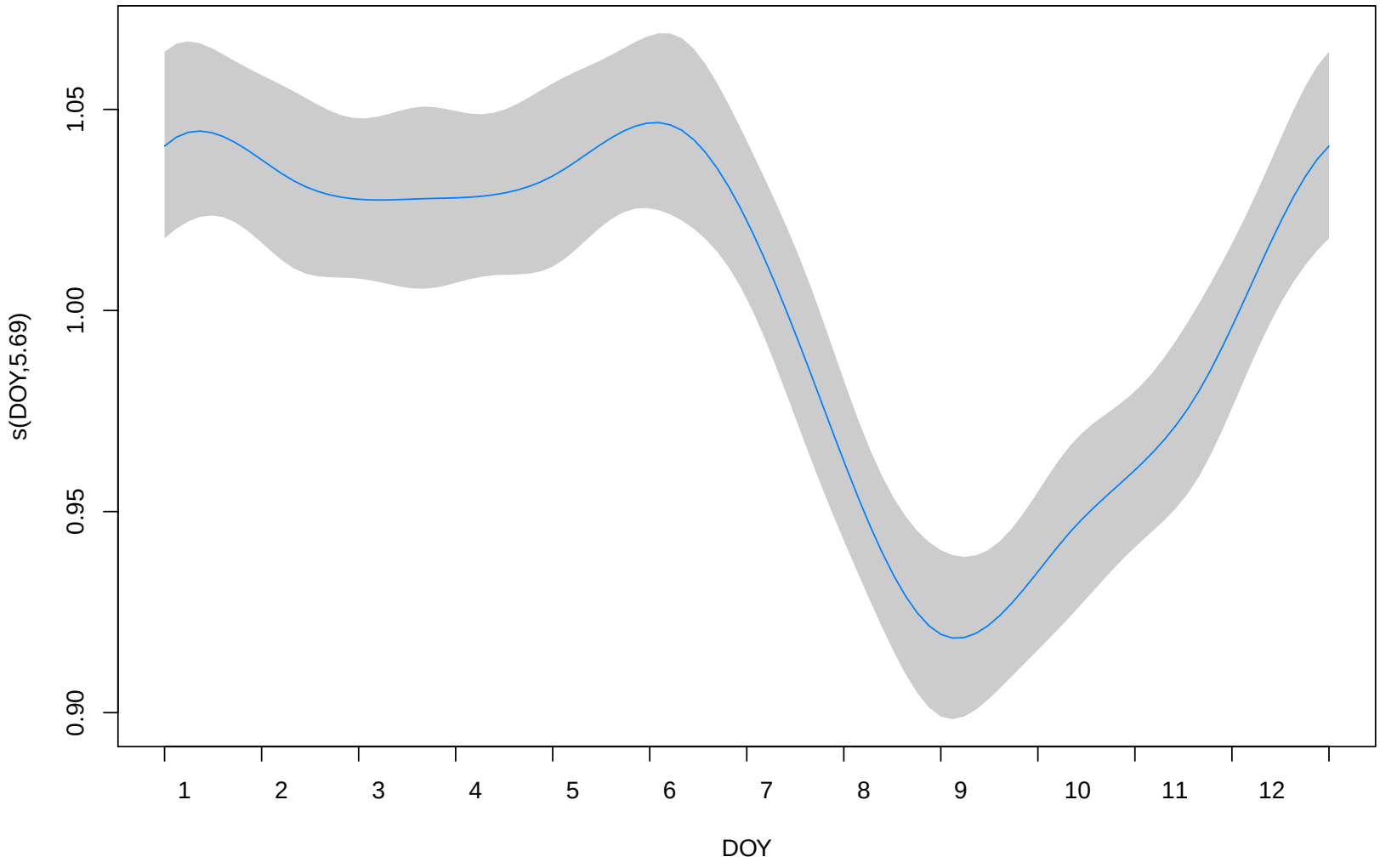
```
par( mfrow = c( 1, 2 ) )
plot( fit, select = 9, scale = 0, rug = FALSE, trans = exp, shade = TRUE, col = trellis.par.get()$superpose.line$col[1] )
plot( fit, select = 10, scale = 0, rug = FALSE, trans = exp, shade = TRUE, col = trellis.par.get()$superpose.line$col[1] )
```





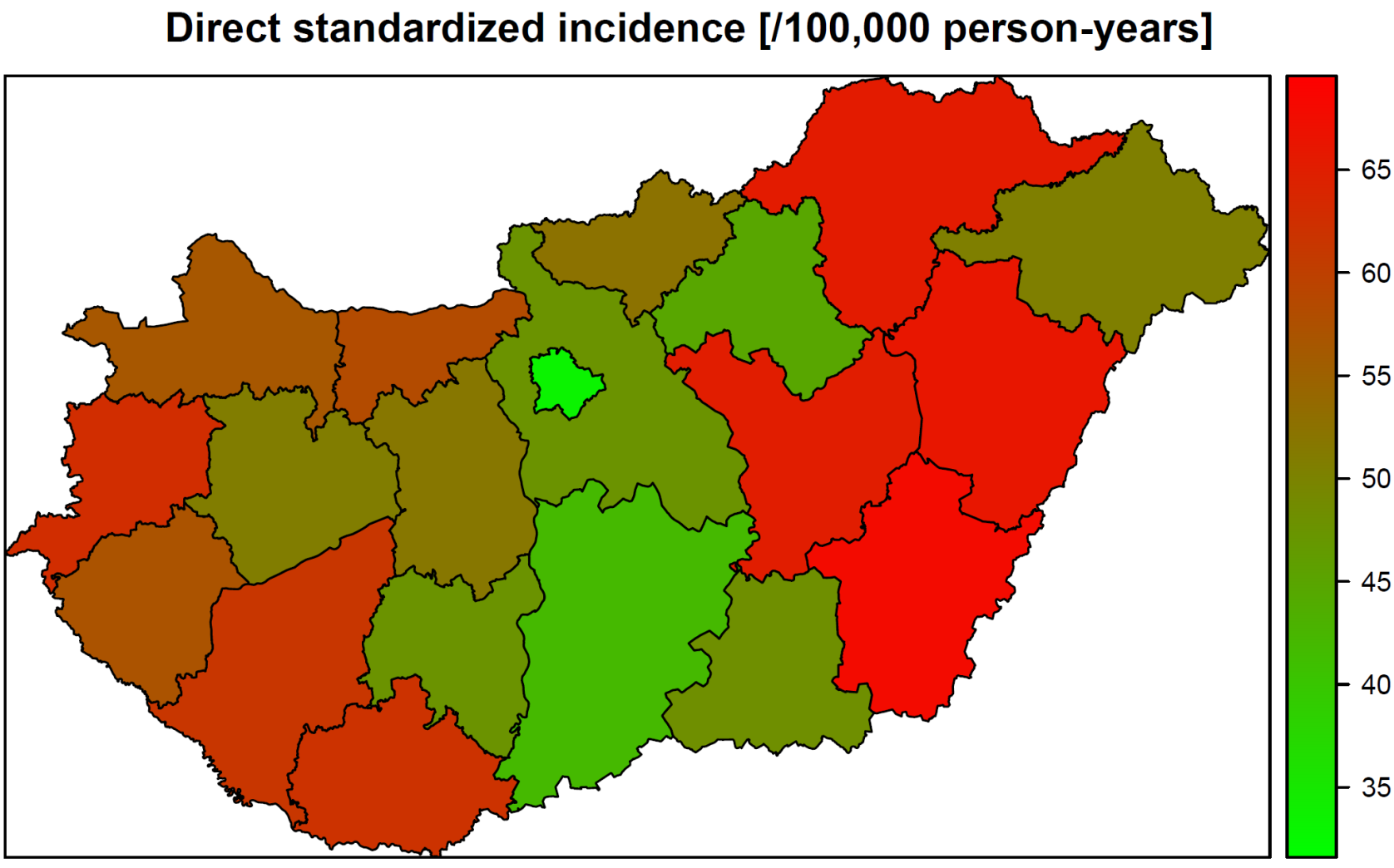
Seasonal pattern

```
plot( fit, select = 8, scale = 0, rug = FALSE, trans = exp, shade = TRUE, col = trellis.par.get()$superpose.line$col[1], xaxt = "n" )
axis( 1, at = seq( 0, 1, 1/12 ), labels = c( 1:12, NA ), hadj = -1 )
```

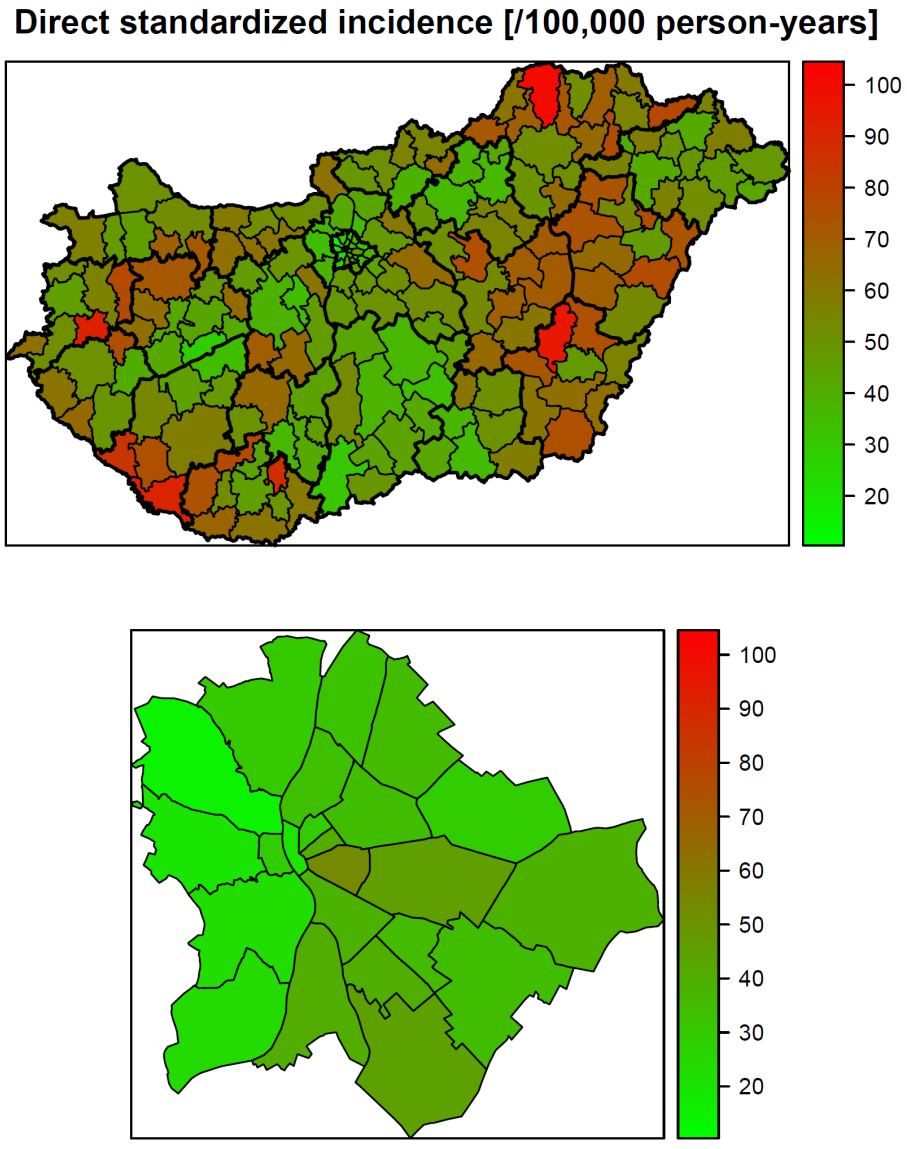


Further questions

Spatial effects



Spatial effects (at another resolution)



Further questions

- Other models, such as mixed models (smoothing)
- Borrowing strength (such as CAR in space)
- Multilevel/hierachical models
- Typically Bayesian estimation
- Spatiotemporal models: the question of time $\times$ space interaction

Thanks for your attention!